A Fast Neural-Dynamical Approach to Scale-Invariant Object Detection

Kasim Terzić, David Lobato, Mário Saleiro and J.M.H. du Buf

Vision Laboratory (ISR-LARSyS), University of the Algarve, Faro, Portugal
{kterzic|dlobato|masaleiro|dubuf}@ualg.pt

Abstract. We present a biologically-inspired method for object detection which is capable of online and one-shot learning of object appearance. We use a computationally efficient model of V1 keypoints to select object parts with the highest information content and model their surroundings by a simple binary descriptor based on responses of cortical cells. We feed these features into a dynamical neural network which binds compatible features together by employing a Bayesian criterion and a set of previously observed object views. We demonstrate the feasibility of our algorithm for cognitive robotic scenarios by evaluating detection performance on a dataset of common household items.

1 Introduction

Reliable detection of objects in complex scenes remains one of the most challenging problems in Computer Vision, despite decades of concentrated effort. Object detection in Cognitive Robotics scenarios imposes further constraints such as real-time performance yet often with limited processing power, so efficient algorithms are needed.

In this paper, we present a fast neural approach to object detection based on cortical keypoints and neural dynamics, which can detect objects from more than 30 classes in real time. The biological foundation of our algorithm is particularly interesting for cognitive robotics based on human vision. We evaluate detection performance on a robotic vision dataset.

1.1 Related Work

Many modern object recognition algorithms begin by a keypoint extraction step to reduce the computational complexity and to discard regions which do not contain useful information. A number of keypoint detectors for extracting points of interest in images are available in the literature [1–3]. In biological vision, retinal input enters area V1 via the Lateral Geniculate Nucleus, and is then processed by layers of so-called simple, complex and end-stopped cells. Simple cells are usually modelled by complex Gabor filters with phases in quadrature, and complex cells by the modulus of the complex response. Simple and complex cells roughly correspond to edge-detectors in Computer Vision. End-stopped cells respond to line terminations, corners, line crossings and blobs, and can thus be

seen as general-purpose keypoint detectors. We base our method on fast V1 keypoints from [4], because they exhibit excellent repeatability and are biologically plausible, which makes them useful for object localisation and recognition.

There are many biologically inspired methods for object detection and recognition. Most of these are based on the Neocognitron and HMAX models, or convolutional neural networks. The Neocognitron architecture [5], originally developed for character recognition, has been successfully applied to object and face recognition [6]. Cortical simple and complex cells form the basis of the HMAX model and its derivatives [7], which alternate pooling and maximum layers to extract features of increasing complexity. However, HMAX requires an external classifier (usually an SVM) for final classification, so it is primarily a feature extraction method. Recently, deep convolutional networks have demonstrated excellent performance on a number of classification tasks, but at a considerable cost in terms of complexity and learning time [8]. The only object recognition algorithm based on neural dynamics known to us was proposed by Faubel and Schöner [9], which jointly estimates object pose and class using dynamic fields.

Concerning object detection and localisation in complex images, like in robotic scenarios, there are several main approaches: (i) sliding windows which apply a classifier at every image position and every scale [10], (ii) salience extraction followed by sequential classification [11], and (iii) voting schemes such as Generalised Hough Transform [12]. Sliding windows are computationally inefficient, while salience operators are typically based on general measures of complexity without object-specific knowledge and therefore do not reliably indicate complete objects. In contrast, our approach is based on voting and grouping, and it can be shown to maximise a Bayesian similarity criterion.

2 Method

2.1 Cortical Keypoints and Binary Descriptors

We begin by applying the fast V1 model from [4]. Given an input image, we compute the cell responses and represent them as sets of neural fields. Responses of simple cells are $R_{\lambda,\theta}$, those of complex cells are $C_{\lambda,\theta}$, and those of double-stopped cells are $D_{\lambda,\theta}$, where λ is the spatial wavelength of the Gabor filters representing simple cells, and θ their orientation. Peaks in the keypoint field $K_\lambda = \sum_\theta D_{\lambda,\theta}$ represent points in the image with high information content.

At each local maximum, we extract a binary keypoint descriptor. The descriptor is represented as a stack of neural maps B^n_λ, where $n \in 1, \ldots, N$ is the dimensionality of the descriptor. Each B^n_λ represents a comparison between the responses of two complex cells within the receptive field of the keypoint, whose size is equal λ:

$$B^n_\lambda = \mathrm{sgn}*(C^n_\lambda - C^{\mathrm{centre}}_\lambda), \tag{1}$$

where $C^{\mathrm{centre}}_\lambda$ is the complex cell in the middle of the receptive field, and $\mathrm{sgn}*(x)$ is 0 if $x \leq 0$ and 1 otherwise. Complex cells C^n_λ are sampled in concentric circles around the keypoint centre.

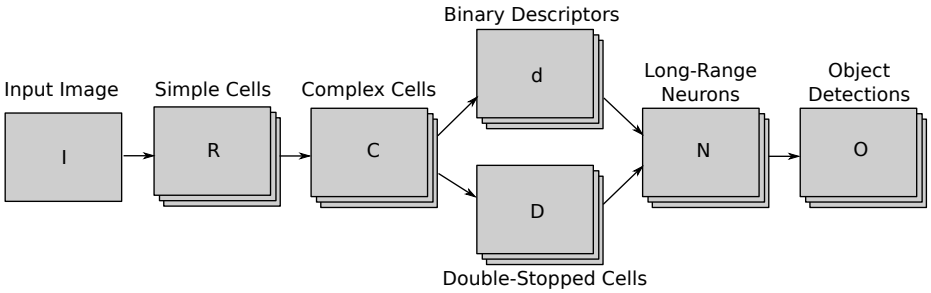

Fig. 1. Overview of the detection process. An input image I is processed by a set of retinotopic neural maps shown in the illustration as stacks of images. We begin by computing V1 responses: simple cells R, complex cells C, and double-stopped cells D, which represent keypoint activations. A local descriptor d is computed for each image location. A set of object-specific cells N responds to keypoints which are similar to those observed during training, and a set of grouping cells combines them into object heat maps O. These are fed into a stack of dynamic fields, which perform non-maximum suppression and pick the best hypothesis at each location.

2.2 Neural Object Detection Model

During training, localised objects are presented to the system and their descriptors d_i are extracted at keypoint locations x_i. For each class $c \in C$, we learn a set of neurons $N_i^{c,s}$ *at the object centre*, where each $N_i^{c,s}$ is associated with a keypoint descriptor *at keypoint location* x_i and s is the scale of the corresponding keypoint. Essentially, each neuron $N_i^{c,s}$ at the object centre has a long-range connection to its actual receptive field at x_i. During testing, keypoints and descriptors are extracted in the same way, but the the weights have been learnt such that the output of $N_i^{c,s}$ is the Hamming distance between a descriptor d_i^{test} from a novel object and a descriptor d_i^{train} observed during training. Therefore, $\sum_i N_i^{c,s}$ evaluates to zero if the system is shown one of the training objects from class c, and it grows large for very different objects.

A codebook of typical features could be learned from $N_i^{c,s}$ using a Self Organising Map, but at the moment we learn a prototype for each keypoint descriptor observed in a training image. We then threshold:

$$\hat{N}_i^{c,s} = [N_j^{c',s} - N_i^{c,s}]^+, \tag{2}$$

with $j \neq i$ and $[\cdot]^+$ represents suppression of negative values. The introduction of the first term ensures that only neurons corresponding to small distances are activated, because large distances are not reliable for probability density estimation [13].

$\hat{N}_i^{c,s}$ are duplicated at all positions of the (subsampled) visual field. They are also duplicated at several scales by scaling the descriptor offset x_i and keypoint scale s by the same factor, resulting in a multi-scale detection framework.

If there is a keypoint in the neuron's receptive field, it is activated and its output is proportional to the thresholded Hamming distance between the observed

descriptor and the training descriptor to which it was tuned. At any given position, a number of neurons $\hat{N}_i^c$ may fire for each class $c \in C$, and the largest of the active neurons are selected and the others are inhibited.

We now define a spatial map of neurons which counts the total accumulated Hamming distance between the observed descriptors and the ones expected by the object model of each class $c \in C$ and convolve it with a circular summing kernel K:

$$M_c(x, y) = \sum_{i,s} \hat{N}_i^{c,s}(x, y) \ , \tag{3}$$

$$O_c = M_c * K \ . \tag{4}$$

We assume that two objects of the same class cannot coexist at the same location in an image, so we sum over scales. This results in only one object map O_c per class.

The convolution with a circular kernel ensures that only features close to the expected position are counted towards a detection of an object, because localised features improve detection [14]. The radius of the kernel represents the maximum acceptable location error of each object part. Object detection now amounts to finding peaks in every O_C and picking the strongest peak at each position.

It can be shown that O_C actually represents at every pixel the logarithm of class likelihood conditioned on observed evidence, when using a nearest neighbour approximation of a naive Bayes classifier [15]. Selecting the class with the highest likelihood thus approximates a Maximum Likelihood classifier, which becomes a Maximum a Posteriori classifier if the prior probabilities of objects are known.

2.3 Winner Selection Using Neural Dynamics

The MAP detection model presented in this work forces a *winner takes all* decision whenever there are two competing detections at the same location. Conversely, as long as the estimate of the object likelihood is valid, picking the strongest hypothesis is equivalent to a local MAP decision between available hypotheses. We achieve this by modelling each object detection map from Eqn 4 as a dynamic field and using two inhibition schemes to force local decisions at peak locations [9].

The first inhibition scheme is global and it is applied to each field separately. It filters out the noise inherent to neural fields. In addition to global inhibition, the resting level of a field is a negative value. It acts as an activation threshold for each object class, thus removing weak and unreliable detections with little support caused by feature noise, i.e. illumination changes, occlusion, etc.

The second inhibition scheme is modelled as field interactions. Each object class is represented by a separate field and they inhibit each other: strong peaks in one field will inhibit smaller peaks at the same location in other fields. It pushes them below the detection threshold, thus forcing a winner-takes-all decision. This ensures that only one object can be detected at any one image location.

2.4 Implementation

A neuron $\hat{N}_i^{c,s}$ is only active if there is a keypoint at scale s located at offset x_i whose descriptor is similar to the one expected by $\hat{N}_i^{c,s}$. This means that the vast majority of all neurons are not active. We exploit this fact in order to improve speed. After extracting keypoints and their descriptors as described, we make each keypoint "vote" for an object centre. We do this by an efficient nearest-neighbour lookup among all descriptors learned during training and finding the k nearest neighbours. Among these, we pick the nearest descriptor for each class, and scale the offset x_i associated with this descriptor by the keypoint scale s. We then activate the neuron $\hat{N}_i^{c,s}$ corresponding to the correct class c, scale s and the scaled offset x_i/s. The distance to neighbour $k+1$ is used to estimate the value of $N_j^{c',s}$ in Eqn 2, as suggested in [13].

The results of this alternative formulation are mathematically equivalent to using a full neural network implemention of our algorithm, but it is orders of magnitude faster and therefore usable for real-time scenarios.

Field-based dynamics were implemented using the CEDAR framework [16]. The early stages of our algorithm are implemented in C++ and OpenCV as a plug-in for CEDAR. Feature extraction and Eqns 3 and 4 were implemented on top of the public keypoint implementation from [4].

3 Evaluation

We evaluate our algorithm on the challenging IIIA30 dataset developed for robot localisation [17]. It consists of cluttered indoor images containing objects from 29 classes, with large scale and pose variance. The objects are annotated using labelled bounding boxes. As per Computer Vision convention, a detection is considered correct if it overlaps with a ground truth annotation of the same class by more than 50% (intersection over union). We compare against two standard methods used in [18]: SIFT keypoints followed by RANSAC grouping (the "classic" SIFT approach), and a Bag-of-Features method based on Vocabulary Trees built on top of SIFT descriptors. We applied our method using two descriptor types: the computational SIFT descriptor, and the biological binary descriptor based on responses of complex cells, introduced in Sec. 2.1. For brevity, we refer to our method as "NDOD": Neural-Dynamic Object Detection. The results for the standard methods were taken from [17].

Table 1 shows a summary of the results, averaged over all classes. The first row shows the best reported F1-score, averaged over all classes. The second and third rows show average recall and precision. Both values were measured separately for each class at the best F1 score for that particular class, then averaged. The fourth row shows mean Average Precision (the area under the precision-recall curve), where available. The last row shows how many classes each detector failed to detect (both recall and precision are zero). It can be seen that our full biological model compares well with the state of the art in computational vision, but does not yet match the classic SIFT approach. However, a combination

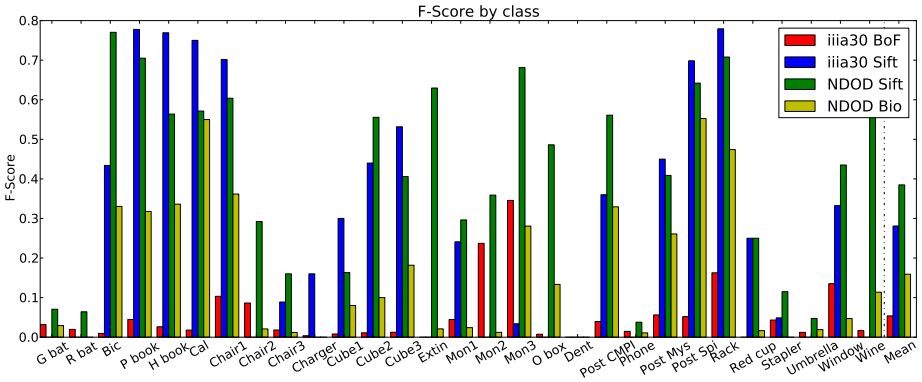

Fig. 2. Best reported F1-score for the 30 object classes from the IIIA30 dataset, compared with two state of the art methods built on SIFT descriptors. Our fully biological model achieves good performance on most classes, and the combination of our model and the SIFT descriptor outperforms all other methods.

	classic SIFT	SIFT+BoF	NDOD+Bio	NDOD+SIFT
Average Best F1 Score	0.281	0.054	0.159	**0.385**
Recall @ Best F1	0.260	**0.408**	0.126	0.346
Precision @ Best F1	0.372	0.032	0.301	**0.497**
Average Precision	n/a	n/a	0.076	**0.217**
% of classes failed	10	3	4	**2**

Table 1. Common performance measures on IIIA30, averaged over all classes (see text). Our fully biological model using a cortical descriptor outperforms the Bag of Features method. It outperforms the SIFT method on many difficult classes, but is weaker on average. Our method combined with the SIFT descriptor outperforms all other methods.

of our neural object detection method and the SIFT descriptor outperforms all other methods, suggesting that a more powerful biologically plausible descriptor would significantly boost performance.

Figure 2 shows a more detailed evaluation of all four detectors. We plot the best reported F1 score for each of the 29 classes, as well as the average. The graph clearly shows that the classic SIFT method works well for some types of objects, and consistently fails with others. Both the Bag-of-Features approach and our method are more reliable with difficult classes. It can be seen that, averaged over all classes, the performance of our full biological model falls half-way between the two computational method. Our method combined with the SIFT descriptor significantly outperforms all other methods. Figure 3 shows some detections on images from the IIIA30 dataset.

Fig. 3. Some detections from the IIIA30 dataset obtained by our method using the SIFT descriptor. We obtain high precision despite blurred images and a cluttered environment.

4 Conclusions

We have presented a real-time object detection mechanism based on cortical keypoints and neural dynamics. Our algorithm performs well on a standard dataset of household objects. To the best of our knowledge, this is the first neural object detection based on dynamic fields, and it significantly advances the state of the art in this field.

While the neural detection model is efficient and works well together with the SIFT descriptor, results show that our current biological image descriptor is holding back the performance of the complete biological model. Luckily, current research in binary image descriptors is often biologically motivated [19, 20], so we expect significant progress in this area. We are currently looking into learning a powerful binary descriptor based on cortical cells, which should replace the hand-crafted one presented in this work.

Our current work focuses on using visual landmarks detected by our algorithm as localisation cues for cognitive robots, leading towards a semantic SLAM implementation.

Acknowledgements This work was supported by the EU under the grant ICT-2009.2.1-270247 *NeuralDynamics* and the Portuguese FCT under the grant PEst-OE/EEI/LA0009/2011.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). CVIU **110** (2008) 346–359
3. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65** (2005) 43–72
4. Terzić, K., Rodrigues, J., du Buf, J.: Fast cortical keypoints for real-time object recognition. In: ICIP, Melbourne (2013) 3372–3376
5. Fukushima, K.: Neocognitron for handwritten digit recognition. Neurocomputing **51** (2003) 161–180
6. Do Huu, N., Paquier, W., Chatila, R.: Combining structural descriptions and image-based representations for image, object, and scene recognition. In: IJCAI. (2005) 1452–1457
7. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Object recognition with cortex-like mechanisms. IEEE T-PAMI **29** (2007) 411–426
8. Schmidhuber, J.: Multi-column deep neural networks for image classification. In: CVPR. (2012) 3642–3649
9. Faubel, C., Schöner, G.: A neuro-dynamic architecture for one shot learning of objects that uses both bottom-up recognition and top-down prediction. In: IROS, IEEE Press (2009) 3162–3169
10. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision **57** (2004) 137–154
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE T-PAMI **20** (1998) 1254–1259
12. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Workshop on Statistical Learning in Computer Vision, ECCV. (2004)
13. McCann, S., Lowe, D.: Local naive bayes nearest neighbor for image classification. In: CVPR, Providence (2012) 3650–3656
14. Mutch, J., Lowe, D.G.: Multiclass Object Recognition with Sparse, Localized Features. In: CVPR. Volume 1., New York (2006) 11–18
15. Terzić, K., du Buf, J.: An efficient naive bayes approach to category-level object detection. In: ICIP, Paris (2014) accepted.
16. Lomp, O., Zibner, S.K.U., Richter, M., Rañó, I., Schöner, G.: A Software Framework for Cognition, Embodiment, Dynamics, and Autonomy in Robotics: Cedar. In: ICANN. (2013) 475–482
17. Ramisa, A.: IIIA30 dataset. http://www.iiia.csic.es/ aramisa/datasets/iiia30.html (2009) [Online. Accessed 30. Apr. 2014.].
18. Ramisa, A.: Localization and Object Recognition for Mobile Robots. PhD thesis, Universitat Autonoma de Barcelona (2009)
19. Leutenegger, S., Chli, M., Siegwart, R.: BRISK: Binary robust invariant scalable keypoints. In: ICCV, Barcelona, IEEE Computer Society (2011) 2548–2555
20. Alahi, A., Ortiz, R., Vandergheynst, P.: FREAK: Fast retina keypoint. In: CVPR, Providence (2012) 510–517